Intrinsic Omnidirectional Video Decomposition

Rong-Kai Xu*
Beijing Institute of Technology

Fang-Lue Zhang†
Victoria University of Wellington

Lei Zhang‡
Beijing Institute of Technology

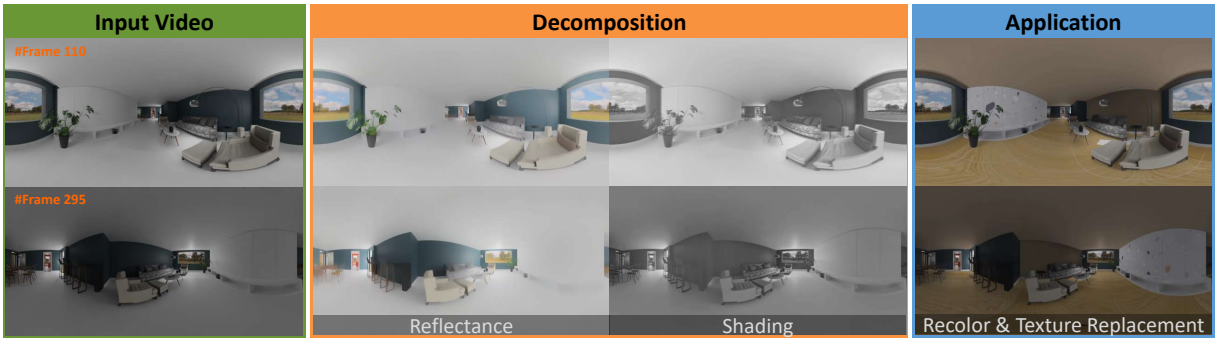

Figure 1: An input omnidirectional video (left) is decomposed into its reflectance and shading components (middle) by the proposed method. The decomposition results can be further used for video manipulation applications (right).

Abstract

Intrinsic decomposition of omnidirectional video is a challenging task. We propose a method that can provide temporally consistent decomposition results. Leveraging the 360-degree scene representation, we maintain the global point cloud to propagate and reuse the similar inter-frame content and establish temporal constraints which elevate the quality of frame-wise decomposition while maintaining inter-frame coherence. By optimizing the proposed objective function, our method achieves a precise separation of reflectance and shading components. Comprehensive experiments demonstrate that our approach outperforms existing intrinsic decomposition methods. Our method also hold promise for various video manipulation applications.

Index Terms: Computing methodologies—Computer graphics—Image manipulation; Human-centered computing—Visualization—Visualization techniques

1 Introduction

Omnidirectional video offers a holistic 360-degree representation of the scene, offering a comprehensive depiction of the environment through a sequence of frames. Intrinsic video decomposition allows for the separation of video into its fundamental reflectance and shading elements for each frame. This offers a detailed set of light-related information, which can be used to enhance the immersive experiences. This decomposition has played a fundamental role in various 2D video manipulation applications, including realistic recoloring and retexturing [10].

Generally, intrinsic video decomposition poses a challenging and inherently ambiguous inverse problem [3], demanding consideration of both spatial and temporal consistency in the decomposition process [7]. The domain of intrinsic video decomposition has seen relatively limited exploration [9, 10], with a predominant focus on enhancing inter-frame consistency. For omnidirectional video, each frame not only encompasses a more comprehensive range of visual information but also exhibits significant correlations across frames. Previous intrinsic decomposition methods have often neglected these distinctive characteristics of omnidirectional content. This oversight results in an underutilization of substantial information that could guide the decomposition process more effectively.

In this study, based on recovered 3D geometric information from the recorded scene, we first attain an initial estimation of the scene's illumination distribution through identifying globally consistent light sources for each frame, which aids in subsequent optimizations for reflectance and shading maps. Moreover, the intricate details of the scene's geometry directly contribute to enforcing uniform reflectance constraints during the final stages of optimizing shading and reflectance maps. Leveraging global light source data and maintaining temporal consistency minimizes redundant computations for similar content, ultimately elevating the contextual significance of the decomposition outcomes.

The main contribution of this work lies in the development of an innovative intrinsic omnidirectional video decomposition method. By leveraging the 360-degree geometric information embedded within omnidirectional video, we introduce a global point cloud representation to capture inter-frame lighting-related content changes. This allows a robust estimation of illumination distributions across frames, enabling the establishment of temporally reflectance and shading constraints. We then proposed an optimization framework to extract the reflectance and shading maps with the established constraints. Our experiment results demonstrate the superior performance of our method, surpassing existing image and video decomposition techniques.

2 Method

Intrinsic video decomposition factors each frame of the input video $\boldsymbol{I} \in \mathbb{R}^{h \times w}$ into a pixel-by-pixel product of reflectance $\boldsymbol{R}$ and shading $\boldsymbol{S}$. Our method utilizes the holistic representation of the scene inherent in omnidirectional videos for intrinsic decomposition. We achieve inter-frame stability through a temporal illumination and reflectance propagation strategy.

We show the overview of our method in Fig. 2. Given an omnidirectional video, we employ Structure-From-Motion to reconstruct the global point cloud of the scene $\boldsymbol{G}$ which contains N points, each point has position, normal and reflectance information. And we

*e-mail: zjlxxrk@gmail.com
†e-mail: fanglue.zhang@vuw.ac.nz
‡e-mail: leizhang@bit.edu.cn

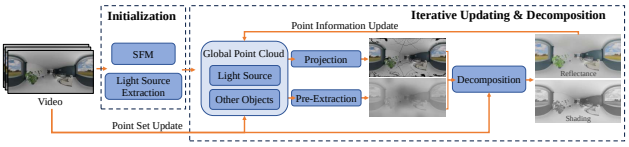

Figure 2: Overview of our method.

extract the light source from several keyframes, which is utilized for initial illumination estimation of each frame. Subsequently, during the decomposition process for each frame, we obtain the initial estimation of the reflectance by projecting the global point cloud to the camera position. And the initial illumination $\boldsymbol{S}_{est}$ is obtained by shading the scene with the extracted light source. Then we introduce temporal reflectance consistency constraints and the temporally smooth illumination constraints and three other constraints to build the decomposition model as follows:

$$\begin{aligned}\arg\min_{\boldsymbol{R},\boldsymbol{S}} \|\omega(\boldsymbol{I}-\boldsymbol{R}\times\boldsymbol{S})\|_2^2+\lambda_{\mathrm{r}}\|\nabla\boldsymbol{R}\|_1+\lambda_{\mathrm{s}}\|\nabla\boldsymbol{S}\|_2^2 \\ +\lambda_{\mathrm{t}}\|\boldsymbol{M_t}(\boldsymbol{R}-\phi(\boldsymbol{G}))\|_2^2+\lambda_{\mathrm{a}}\|\boldsymbol{S}-\boldsymbol{S}_{\mathrm{est}}\|_2^2\end{aligned} \tag{1}$$

where ϕ represents the reflectance obtained by projecting the point cloud to current camera position and the M_t represents the mask of new pixels in current frame. This non-convex optimization problem can be solved by the iteratively reweighted least squares (IRLS) solver to generate the final reflectance and shading of current frame.

During the frame-by-frame decomposition process, the information within the point cloud $\boldsymbol{G}$ undergoes iterative updates. These updates occur both prior to and following the decomposition of each frame. Before processing a frame, it is crucial to account for the potential emergence of new scene points that become visible due to camera motion. These points should be added to the point cloud G. To prevent the inclusion of duplicate scene points at this stage, we employ the projection of the previous 3D point set onto the current frame's camera to verify whether the corresponding 3D point for a pixel in the current frame has been previously observed. In addition, for enhanced decomposition efficiency, we cull 3D points that remain invisible for an extended period from the point cloud. This ensures a consistently stable decomposition speed throughout the entire video. Following the decomposition of each frame, we save a pixel's reflectance value to its corresponding 3D point. This will serve as a reflectance consistency constraint for the decomposition of the next frame at the corresponding position.

Table 1: Quantitative evaluation on the decomposition results of the sequence of *Video1* processed using the method of [1].

Method	sMSE↓	sSMSE↓	sLMSE↓	SSIM↑
Das et al. [2]	**0.0243**	0.0217	0.0225	0.7908
Luo et al. [6]	0.0478	0.0241	0.0229	0.7360
Li et al. [5]	0.0708	0.0254	0.0237	0.7270
Zhu et al. [11]	0.0384	0.0327	0.0275	0.7410
Li et al. [4]	0.0977	0.0450	0.0416	0.7208
Ours	0.0244	**0.0198**	**0.0196**	**0.8381**

3 Evaluation

We conduct experiments to verify the effectiveness of our method. Due to the absence of indoor omnidirectional video datasets suitable for intrinsic decomposition quantitative analysis, we generated synthetic clips using rendering software, each contains 300 frames. Our method takes about 20s to decompose a 512×1024 frame. And due to our point cloud update strategy, the point cloud G only contains $N \approx 3 \times h \times w$ points on average.

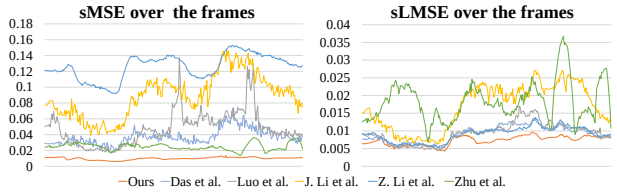

Figure 3: Statistics of sMSE and sLMSE over the frames of *Video2*. Our method shows better stability and lower errors.

As shown in Table 1, our approach outperforms other methods across all four metrics. To better assess video stability, we follow the approach of Meka et al. [8] and plot the per-frame error. Fig. 3 shows that our method exhibits a lower curve, indicating a lower average error, and maintains a smoother curve compared to both image and video decomposition methods, showcasing superior stability.

4 Discussion and Conclusion

We have introduced a novel method for intrinsic omnidirectional video decomposition. Comprehensive evaluations demonstrate that our approach surpasses current state-of-the-art techniques, highlighting its potential for a wide range of video editing applications. However, there are some limitations that our method isn't efficient enough for real-time decomposition and only takes the static scene into consideration. We aim to explore more efficient and reliable approaches to solve these problems in our future work.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2021QY1503. Fang-Lue Zhang was supported by the Marsden Fund under grant MFP-20-VUW-180.

References

[1] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister. Blind video temporal consistency. *ACM Transactions on Graphics*, 34(6), 2015.

[2] P. Das, S. Karaoglu, and T. Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of CVPR*, pp. 19790–19799, 2022.

[3] E. Garces, C. Rodriguez-Pardo, D. Casas, and J. Lopez-Moreno. A survey on intrinsic images: Delving deep into lambert and beyond. *International Journal of Computer Vision*, 130(3):836–868, 2022.

[4] J. Li, H. Li, and Y. Matsushita. Lighting, reflectance and geometry estimation from 360 panoramic stereo. In *Proceedings of CVPR*, pp. 10586–10595, 2021.

[5] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of CVPR*, pp. 2475–2484, 2020.

[6] J. Luo, Z. Huang, Y. Li, X. Zhou, G. Zhang, and H. Bao. Niid-net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3434–3445, 2020.

[7] A. Meka, G. Fox, M. Zollhöfer, C. Richardt, and C. Theobalt. Live user-guided intrinsic video for static scenes. *IEEE Transactions on Visualization and Computer Graphics*, 23(11):2447–2454, 2017.

[8] A. Meka, M. Shafiei, M. Zollhöfer, C. Richardt, and C. Theobalt. Real-time global illumination decomposition of videos. *ACM Transactions on Graphics*, 40(3):1–16, 2021.

[9] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live intrinsic video. *ACM Transactions on Graphics*, 35(4):1–14, 2016.

[10] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics.*, 33(4), jul 2014.

[11] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi, R. Wang, H. Bao, J. Zheng, and R. Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia*, pp. 1–8, 2022.